**Big Security Data – Project Milestone 2**

Group #13: Kindell Holst, Denesha Record, William McGuire

April 18, 2025

**Dataset Identification and Justification**

A huge threat discussed in the first milestone to the insurance industry and the world at large is the use of social engineering to get access to sensitive and private information. The number one form of social engineering in the present day is phishing. Phishing occurs when messages are sent to employees of a company, or third parties related to the company that contain links to supposed normal sites. These sites, however, are not normal. They are a way to get credential information and/or private information from the victim, allowing for data access for malicious actors. As a company, guards need to be put in place for this, to ensure that phishing attacks are snuffed out before the damage is done.

The first data set we will be using is a phishing websites data set that was put together by Rami M. Mohammed and Lee McCluskey at The University of Huddersfield in Huddersfield, UK. It was made using data stores like PhishTank archive, MillerSmiles archive, and also Googles searching operators. This data set has a very defined purpose, to create a machine learning model that can identify phishing websites with simple domain, website, and url information.

Looking into virustotal api.

**Collection and Parsing Strategies**

The phishing websites dataset was collected on UC Irvine Machine Learning Repository. This repository includes 677 labelled public datasets to be used for machine learning projects. It also comes with a word document discussing the different logical reasons for features being important to identifying phishing websites. The data set is stored in an ARFF file. ARFF means Attribute-Relation File Format, and they are used for machine learning data sets on WEKA, an open-source machine learning software. As we are not using WEKA for our analysis, we converted the file into a csv format.

**Data Summary**

**Group Roles and Signatures**